

In the n -gram or text classification methods we've seen so far, words are just strings – that's not very satisfactory!

What do we want from a theory of word meaning?

We can get some ideas from *lexical semantics* – the linguistic study of word meaning.

mouse

mouse deer

mouse-eared bat

mouse lemur

mouse mat

mouse mats

mouse opossum

mouse pad

mouse potato

mouse-tailed bat

mousebird

mousemat

mouseover

mouser

mousetrap

mousey

mouse-like

< | >

Dictionary

17 found

A | A

Q mouse

×

All Dictionary Thesaurus Apple Wikipedia

mouse

noun (*plural mice* | mīs |) | mous |

1 a small rodent that typically has a pointed snout, relatively large ears and eyes, and a long tail.

Family Muridae: many genera and numerous species. Also, some species in the families Heteromyidae, Zapodidae, and Muscardinidae

- (in general use) any small mammal similar to a mouse, such as a shrew or vole.
- a shy, timid, and quiet person: *Jane may be a bit of a mouse, but she is very nosy.*
- a dull light brown color reminiscent of a mouse's fur: *her blonde hair dulled to mouse.*

2 (*plural mice* or *plural mouses*) a small handheld device that is dragged across a flat surface to move the cursor on a computer screen, typically having buttons that are pressed to control functions: *copy the file with a click of the mouse* | [*as modifier*] : *the right mouse button.*

Lemma

mouse

mouse deer
mouse-eared bat
mouse lemur
mouse mat
mouse mats
mouse opossum
mouse pad
mouse potato
mouse-tailed bat
mousebird
mousemat
mouseover
mouser
mousetrap
mousey
mouse-like

Dictionary
17 found

All Dictionary Thesaurus Apple Wikipedia

mouse

noun (*plural* **mice** | mīs |) | mous |

1 a small rodent that typically has a pointed snout, relatively large ears and eyes, and a long tail.

Family Muridae: many genera and numerous species. Also, some species in the families Heteromyidae, Zapodidae, and Muscardinidae

- (in general use) any small mammal similar to a mouse, such as a shrew or vole.
- a shy, timid, and quiet person: *Jane may be a bit of a mouse, but she is very nosy.*
- a dull light brown color reminiscent of a mouse's fur: *her blonde hair dulled to mouse.*

2 (*plural* **mice** or *plural* **mouses**) a small handheld device that is dragged across a flat surface to move the cursor on a computer screen, typically having buttons that are pressed to control functions: *copy the file with a click of the mouse* | [*as modifier*] : *the right mouse button.*

Lemma

Senses

mouse

mouse deer

mouse-eared bat

mouse lemur

mouse mat

mouse mats

mouse opossum

mouse pad

mouse potato

mouse-tailed bat

mousebird

mousemat

mouseover

mouser

mousetrap

mousey

mouse-like

Dictionary
17 found

All Dictionary Thesaurus Apple Wikipedia

mouse

noun (*plural mice* | mīs |) | mous |

1 a small rodent that typically has a pointed snout, relatively large ears and eyes, and a long tail.

Family Muridae: many genera and numerous species. Also, some species in the families Heteromyidae, Zapodidae, and Muscardinidae

- (in general use) any small mammal similar to a mouse, such as a shrew or vole.
- a shy, timid, and quiet person: *Jane may be a bit of a mouse, but she is very nosy.*
- a dull light brown color reminiscent of a mouse's fur: *her blonde hair dulled to mouse.*

2 (*plural mice* or *plural mouses*) a small handheld device that is dragged across a flat surface to move the cursor on a computer screen, typically having buttons that are pressed to control functions: *copy the file with a click of the mouse* | [*as modifier*] : *the right mouse button.*

<

|

>

☰

en-word.net/view/lemma/mouse

↺

📄



Open English
Wordnet

Enter a word or identifier

Options ▼

Nouns

(n)

mouse

any of numerous small rodents typically resembling diminutive rats having pointed snouts and small ears on elongated bodies with slender usually hairless tails

MORE ▶

(n)

shiner, black eye, mouse

a swollen bruise caused by a blow to the eye

MORE ▶

(n)

mouse

person who is quiet or timid

MORE ▶

(n)

mouse, computer mouse

a hand-operated electronic device that controls the coordinates of a cursor on your computer screen as you move it around on a pad; on the bottom of the device is a ball that rolls on the surface of the pad "a mouse takes much more room than a trackball"

MORE ▶

Verbs

(v)

sneak, mouse, creep, pussyfoot

to go stealthily or furtively "…stead of sneaking around spying on the neighbor's house"

MORE ▶

(v)

mouse

manipulate the mouse of a computer

MORE ▶

Download As:

[JSON](#)

[RDF/XML](#)

[Turtle](#)

[XML](#)

Think–pair–share

What relation holds between the pairs of words in each of the word banks below?

big / large
oculist / eye-doctor
car / automobile
water / H₂O
draft / draught

yes / no
dark / light
hot / cold
up / down
happy / sad

cat / dog
cardiologist /
pulmonologist
car / bus
sheep / goat
glass / mug

coffee / cup
waiter / menu
farm / cow
house / roof
theater / actor

Think–pair–share

big / large
oculist / eye-doctor
car / automobile
water / H₂O
draft / draught

yes / no
dark / light
hot / cold
up / down
happy / sad

cat / dog
cardiologist /
pulmonologist
car / bus
sheep / goat
glass / mug

coffee / cup
waiter / menu
farm / cow
house / roof
theater / actor

Think–pair–share

Synonymy

big / large
oculist / eye-doctor
car / automobile
water / H₂O
draft / draught

yes / no
dark / light
hot / cold
up / down
happy / sad

cat / dog
cardiologist /
pulmonologist
car / bus
sheep / goat
glass / mug

coffee / cup
waiter / menu
farm / cow
house / roof
theater / actor

Think–pair–share

Synonymy

big / large

oculist / eye-doctor

car / automobile

water / H₂O

draft / draught

my big sister ≠ my large sister

dark / light

hot / cold

up / down

happy / sad

cat / dog

*cardiologist /
pulmonologist*

car / bus

sheep / goat

glass / mug

coffee / cup

waiter / menu

farm / cow

house / roof

theater / actor

Think–pair–share

Synonymy

big / large

oculist / eye-doctor

car / automobile

water / H₂O

draft / draught

my big sister ≠ my large sister

dark / light

hot / cold

Would you use H₂O in a surfing guide?

happy / sad

cat / dog

*cardiologist /
pulmonologist*

car / bus

sheep / goat

glass / mug

coffee / cup

waiter / menu

farm / cow

house / roof

theater / actor

LA JUSTESSE
DE LA
LANGUE FRANÇOISE,
OU
LES DIFFERENTES SIGNIFICATIONS
DES MOTS QUI PASSENT
POUR
SYNONIMES.

Par M. l'Abbé GIRARD C. D. M. D. D. B.



A PARIS,
Chez LAURENT D'HOURY, Imprimeur-
Libraire, au bas de la rue de la Harpe, vis-
à vis la rue S. Severin, au Saint-Esprit.

M. DCC. XVIII.

Avec Approbation & Privilège du Roy.

je ne crois pas qu'il y ait de
mot synonyme dans aucune
Langue.

*[I do not believe that there is a synonymous word
in any language]*

Think–pair–share

Synonymy

big / large

oculist / eye-doctor

car / automobile

water / H₂O

draft / draught

yes / no

dark / light

hot / cold

up / down

happy / sad

cat / dog

*cardiologist /
pulmonologist*

car / bus

sheep / goat

glass / mug

coffee / cup

waiter / menu

farm / cow

house / roof

theater / actor

Think–pair–share

Synonymy

big / large
oculist / eye-doctor
car / automobile
water / H₂O
draft / draught

Antonymy

yes / no
dark / light
hot / cold
up / down
happy / sad

cat / dog
cardiologist /
pulmonologist
car / bus
sheep / goat
glass / mug

coffee / cup
waiter / menu
farm / cow
house / roof
theater / actor

Think–pair–share

Synonymy

big / large
oculist / eye-doctor
car / automobile
water / H₂O
draft / draught

Antonymy

yes / no
dark / light
hot / cold
up / down
happy / sad

cat / dog
cardiologist /
pulmonologist
car / bus
sheep / goat
glass / mug

coffee / cup
waiter / menu
farm / cow
house / roof
theater / actor

Think–pair–share

Synonymy

big / large
oculist / eye-doctor
car / automobile
water / H₂O
draft / draught

Antonymy

yes / no
dark / light
hot / cold
up / down
happy / sad

Similarity

cat / dog
cardiologist /
pulmonologist
car / bus
sheep / goat
glass / mug

coffee / cup
waiter / menu
farm / cow
house / roof
theater / actor

Think–pair–share

Synonymy

big / large
oculist / eye-doctor
car / automobile
water / H₂O
draft / draught

Antonymy

yes / no
dark / light
hot / cold
up / down
happy / sad

Similarity

cat / dog
cardiologist /
pulmonologist
car / bus
sheep / goat
glass / mug

coffee / cup
waiter / menu
farm / cow
house / roof
theater / actor

Think–pair–share

Synonymy

big / large
oculist / eye-doctor
car / automobile
water / H₂O
draft / draught

Antonymy

yes / no
dark / light
hot / cold
up / down
happy / sad

Similarity

cat / dog
cardiologist /
pulmonologist
car / bus
sheep / goat
glass / mug

Association

coffee / cup
waiter / menu
farm / cow
house / roof
theater / actor

Semantic field

Words that

cover a particular semantic domain

bear structured relations with each other

Examples:

hospitals *surgeon, scalpel, nurse, anesthetic, hospital*

restaurants *waiter, menu, plate, food, chef*

houses *door, roof, kitchen, family, bed*

Can we build a representation of these kinds of word meanings?

If you haven't heard the word before, how could you learn what *pawpaw* means?

PAWPAW, MOST NEGLECTED AMERICAN FRUIT.

THE season of the pawpaw is here. The pawpaw is the most neglected of American fruits; few people like it, and somehow it is as despised as the mule and as unpopular as the shitepoke. People always laugh when the pawpaw is mentioned, though why, Heaven knows. For it has a rich and luscious flavor—not unlike the custard apple of the Hesperides. It digests easily in the human stomach, and cheers but does not inebriate. Yet humanity has left it alone. It crowns no festal board. It forms the centre of no famous dish. The pawpaw is not pickled, preserved, dried or stored away against any day of want. No pawpaw pie nor pawpaw pudding is known to man; the pawpaw stewed or baked or fried never has appeared upon any housewife's table, however humble it may be. In all the famine that followed the droughts of the early days of the pioneers when want was forever peering around the corner of tomorrow, no one even turned to the pawpaw for preservation. The coffee bean, the sheep sorrel, the pignut, the sorghum plant, the hackberry—a lean and cadaverous refuge—the dandelion and the poke, all helped men stave off the famine and the fever, while the willing pawpaw lingered amiably about,

and no one noticed or cared. It never served, although it was glad to "stand and wait."

Probably the pawpaw's "limitation" comes from the fact that it is hard to eat with any dignity. No woman would dare eat a pawpaw in the presence of her lover; the sight is disgusting to the point of utter disillusion. Byron is said to have hated to see a woman spit. If he could have seen a woman eat a pawpaw, he would have been a different man. "Don Juan" might never have been written. No one can eat a pawpaw who is not willing to revert to barbarism, to jump back to the forest. A pawpaw probably was intended for a feast in the tree crotch, where spitting seeds is easy and natural. A monkey might get away with a pawpaw gracefully, but a man must be willing to surrender his dignity for the joy of the fruit.

So, we have the pawpaw neglected. Some other age may come to the pawpaw and find there the elixir of youth, the nectar of the gods, the fountain of immortality. For surely wise nature has not created so delicious a fruit in vain. Man is not up to the pawpaw yet. He left it when he descended from the trees, and is not worthy yet to have it back.—*Emporia Daily Gazette.*

The New York Times,
19 November 1922

PAWPAW, MOST NEGLECTED AMERICAN FRUIT.

THE season of the pawpaw is here. The pawpaw is the most neglected of American fruits; few people like it, and somehow it is as despised as the mule and as unpopular as the shitepoke. People always laugh when the pawpaw is mentioned, though why, Heaven knows. For it has a rich and luscious flavor—not unlike the custard apple of the Hesperides. It digests easily in the human stomach, and cheers but does not inebriate. Yet humanity has left it alone. It crowns no festal board. It forms the centre of no famous dish. The pawpaw is not pickled, preserved, dried or stored away against any day of want. No pawpaw pie nor pawpaw pudding is known to man; the pawpaw stewed or baked or fried never has appeared upon any housewife's table, however humble it may be. In all the famine that followed the droughts of the early days of the pioneers when want was forever peering around the corner of tomorrow, no one even turned to the pawpaw for preservation. The coffee bean, the sheep sorrel, the pignut, the sorghum plant, the hackberry—a lean and cadaverous refuge—the dandelion and the poke, all helped men stave off the famine and the fever, while the willing pawpaw lingered amiably about,

and no one noticed or cared. It never served, although it was glad to "stand and wait."

Probably the pawpaw's "limitation" comes from the fact that it is hard to eat with any dignity. No woman would dare eat a pawpaw in the presence of her lover; the sight is disgusting to the point of utter disillusion. Byron is said to have hated to see a woman spit. If he could have seen a woman eat a pawpaw, he would have been a different man. "Don Juan" might never have been written. No one can eat a pawpaw who is not willing to revert to barbarism, to jump back to the forest. A pawpaw probably was intended for a feast in the tree crotch, where spitting seeds is easy and natural. A monkey might get away with a pawpaw gracefully, but a man must be willing to surrender his dignity for the joy of the fruit.

So, we have the pawpaw neglected. Some other age may come to the pawpaw and find there the elixir of youth, the nectar of the gods, the fountain of immortality. For surely wise nature has not created so delicious a fruit in vain. Man is not up to the pawpaw yet. He left it when he descended from the trees, and is not worthy yet to have it back.—*Emporia Daily Gazette.*

The New York Times,
19 November 1922

Pawpaw Recommended by U.S. Food Experts, Along With Persimmon, as War Nutrition

By The United Press.

WASHINGTON, April 20—The Department of Agriculture today declared unequivocally that persimmons are "good eating." The same claim was made for pawpaws, Juneberries, buffalo berries, the high-bush cranberry of the Great Plains, the Western sand cherry and elderberries.

Not only are persimmons and the others palatable, the department said, but they merit, in addition to "useful" and "valuable," such warmer adjectives as "excellent," "delicious" and "rich." There is however, a good reason for this official lyricism—the war.

With sugar becoming more scarce by the day, "so much natural sweetness should be appreciated," the department believes, and "in times like these, American housewives will be wise to get acquainted with their native wild fruits."

The department might not have spoken out so boldly in behalf of persimmons, papaws and the rest were the British not doing equally well with "rose hips," the fruit of the English wild rose, consisting of seed-like pods which are rich in vitamins. These are replacing the

oranges and other citrus fruits whose import has been cut off by the war.

Persimmons, according to the agricultural announcement, can do for Americans what "rose hips" are doing for the British. Further, persimmons merit popularity, it is officially stressed, because their ripe fruit "has flesh about as soft as baked apple or baked custard," is useful "in cakes, custards, sherbets and puddings," is "richer in calories than bananas" and is "as rich in Vitamin C as oranges, grapefruit and lemons."

The pawpaw or "Indian banana" is similarly nourishing, being high in calories and protein, and its fruit "is rich, highly perfumed, about as soft as custard."

The Rocky Mountain region's wild Juneberry, very sweet, is "excellent in pie or dessert combined with some sour fruit like rhubarb or sour cherry," and the buffalo berry was described as "very delicious for sauce, preserves, jelly and jam."

The high-bush cranberry makes good jelly, and the Western sand cherry works up well into sauce, pie or jam.

The New York Times,
21 April 1942

Pawpaw Recommended by U.S. Food Experts, Along With Persimmon, as War Nutrition

By The United Press.

WASHINGTON, April 20—The Department of Agriculture today declared unequivocally that persimmons are "good eating." The same claim was made for pawpaws, Juneberries, buffalo berries, the high-bush cranberry of the Great Plains, the Western sand cherry and elderberries.

Not only are persimmons and the others palatable, the department said, but they merit, in addition to "useful" and "valuable," such warmer adjectives as "excellent," "delicious" and "rich." There is however, a good reason for this official lyricism—the war.

With sugar becoming more scarce by the day, "so much natural sweetness should be appreciated," the department believes, and "in times like these, American housewives will be wise to get acquainted with their native wild fruits."

The department might not have spoken out so boldly in behalf of persimmons, papaws and the rest were the British not doing equally well with "rose hips," the fruit of the English wild rose, consisting of seed-like pods which are rich in vitamins. These are replacing the

oranges and other citrus fruits whose import has been cut off by the war.

Persimmons, according to the agricultural announcement, can do for Americans what "rose hips" are doing for the British. Further, persimmons merit popularity, it is officially stressed, because their ripe fruit "has flesh about as soft as baked apple or baked custard," is useful "in cakes, custards, sherbets and puddings," is "richer in calories than bananas" and is "as rich in Vitamin C as oranges, grapefruit and lemons."

The pawpaw or "Indian banana" is similarly nourishing, being high in calories and protein, and its fruit "is rich, highly perfumed, about as soft as custard."

The Rocky Mountain region's wild Juneberry, very sweet, is "excellent in pie or dessert combined with some sour fruit like rhubarb or sour cherry," and the buffalo berry was described as "very delicious for sauce, preserves, jelly and jam."

The high-bush cranberry makes good jelly, and the Western sand cherry works up well into sauce, pie or jam.

The New York Times,
21 April 1942

which contain toxins. Many people also [cook](#) with ripe pawpaws, making [bread](#), [beer](#), [ice cream](#) or this [pawpaw pudding](#) from NYT Cooking.

The New York Times,
19 October 2020

Many people also cook with ripe pawpaws,
making bread, beer, ice cream or this pawpaw pudding from NYT
Cooking.

The New York Times,
19 October 2020



The pawpaw is also pollinated by flies and other insects rather than by honeybees, said Mr. Dain, and it flowers over several weeks instead of all at once, which ensures that fruit isn't lost to the Northeast's spring frosts.

The New York Times,
19 October 2020

This sounds like another sense – the tree rather than the fruit.

“You shall know a word by the company it keeps!”

J. R. Firth, 1957



Cambridge



Amandalynne Paullada

@amandalynneP

...

TIL that when Firth (1957) said, "You shall know a word by the company it keeps!," the word he was talking about was "ass."

The *placing* of a *text* as a constituent in a context of situation contributes to the statement of meaning since situations are set up to recognize *use*. As Wittgenstein says, 'the meaning of words lies in their use.'⁴ The day to day practice of playing language games recognizes customs and rules. It follows that a text in such established usage may contain sentences such as 'Don't be such an ass!', 'You silly ass!', 'What an ass he is!' In these examples, the word *ass* is in familiar and habitual company, commonly collocated with *you silly—, he is a silly—, don't be such an—*. You shall know a word by the company it keeps! One of the meanings of *ass* is its habitual collocation with such other words as those above quoted.⁵ Though Wittgenstein was dealing with another problem, he also recognizes the plain face-value, the physiognomy of words. They look at us!⁶ 'The sentence is composed of the words and that is enough.'

12:04 AM · Oct 15, 2021

What words can appear in these contexts?

Word 1

*drown, bathroom,
shower, fill, fall, lie,
electrocute, toilet,
whirlpool, iron, gin*

Word 2

*eat, fall, pick, slice,
peel, lie, tree, throw,
fruit, pie, bite, crab,
grate*

Word 3

*advocate, overthrow,
establish, citizen, ideal,
representative,
dictatorship, campaign,
bastion, freedom*

Word 4

*spend, enjoy,
remember, last, pass,
end, die, happen,
brighten, relive*

What words can appear in these contexts?

bathtub

*drown, bathroom,
shower, fill, fall, lie,
electrocute, toilet,
whirlpool, iron, gin*

Word 2

*eat, fall, pick, slice,
peel, lie, tree, throw,
fruit, pie, bite, crab,
grate*

Word 3

*advocate, overthrow,
establish, citizen, ideal,
representative,
dictatorship, campaign,
bastion, freedom*

Word 4

*spend, enjoy,
remember, last, pass,
end, die, happen,
brighten, relive*

What words can appear in these contexts?

bathtub

*drown, bathroom,
shower, fill, fall, lie,
electrocute, toilet,
whirlpool, iron, gin*

apple

*eat, fall, pick, slice,
peel, lie, tree, throw,
fruit, pie, bite, crab,
grate*

Word 3

*advocate, overthrow,
establish, citizen, ideal,
representative,
dictatorship, campaign,
bastion, freedom*

Word 4

*spend, enjoy,
remember, last, pass,
end, die, happen,
brighten, relive*

What words can appear in these contexts?

bathtub

*drown, bathroom,
shower, fill, fall, lie,
electrocute, toilet,
whirlpool, iron, gin*

apple

*eat, fall, pick, slice,
peel, lie, tree, throw,
fruit, pie, bite, crab,
grate*

democracy

*advocate, overthrow,
establish, citizen, ideal,
representative,
dictatorship, campaign,
bastion, freedom*

Word 4

*spend, enjoy,
remember, last, pass,
end, die, happen,
brighten, relive*

What words can appear in these contexts?

bathtub

*drown, bathroom,
shower, fill, fall, lie,
electrocute, toilet,
whirlpool, iron, gin*

apple

*eat, fall, pick, slice,
peel, lie, tree, throw,
fruit, pie, bite, crab,
grate*

democracy

*advocate, overthrow,
establish, citizen, ideal,
representative,
dictatorship, campaign,
bastion, freedom*

day

*spend, enjoy,
remember, last, pass,
end, die, happen,
brighten, relive*

The *distributional hypothesis*: Similar words appear in similar contexts.

Therefore, we can measure similarity in meaning as similarity in contexts.

One more:

Word 5

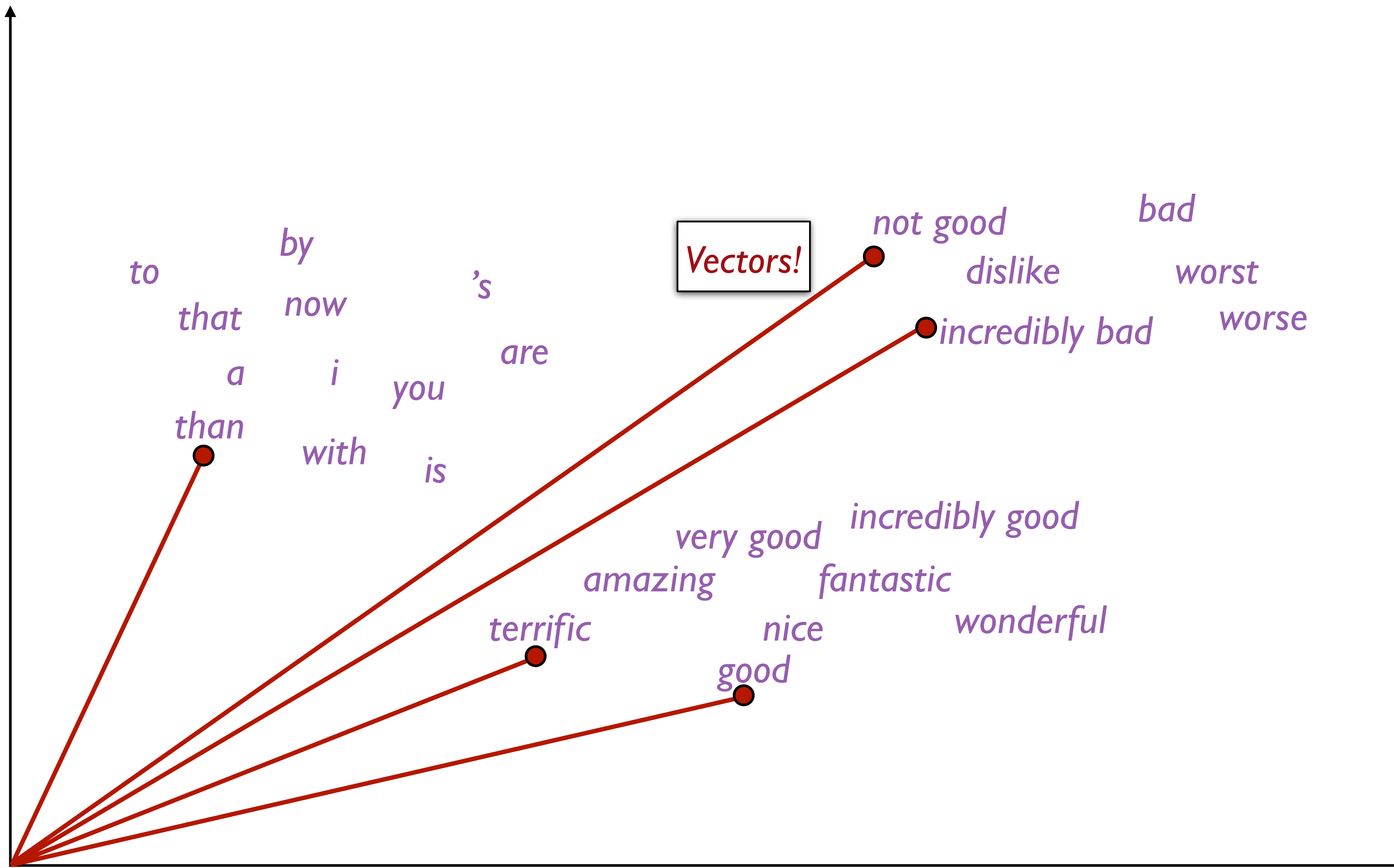
*eat, paint, peel, last,
apple, fruit, juice,
lemon, blue, grow*

One more:

orange

*eat, paint, peel, last,
apple, fruit, juice,
lemon, blue, grow*

Vector semantics instantiates the distributional hypothesis by learning the representations of the meanings of words directly from their distributions in texts.



*Words used in the same context will
cluster nearby in the same space.*



The vector representations of words are called *embeddings*.

In *static embeddings*, each *word type* has a unique embedding.

In *contextual embeddings*, each *word token* has a unique embedding.

This is useful for learning multiple senses.

We'll return to contextual embeddings when we study LLMs.

This week, we'll look at two kinds of static embeddings:

Today we'll look at *simple count embeddings*, where words are represented by the counts of nearby words

Thursday we'll look at *Word2vec*, where representation is created by training a classifier to predict whether a word is likely to appear nearby

Count-based embeddings

Word vectors

	<i>battle</i>	<i>good</i>	<i>fool</i>	<i>wit</i>
<i>As You Like It</i>	1	114	36	20
<i>Twelfth Night</i>				
<i>Julius Caesar</i>				
<i>Henry V</i>				

Document–word matrix

*A document can be represented
as the number of times each
word occurs in it.*

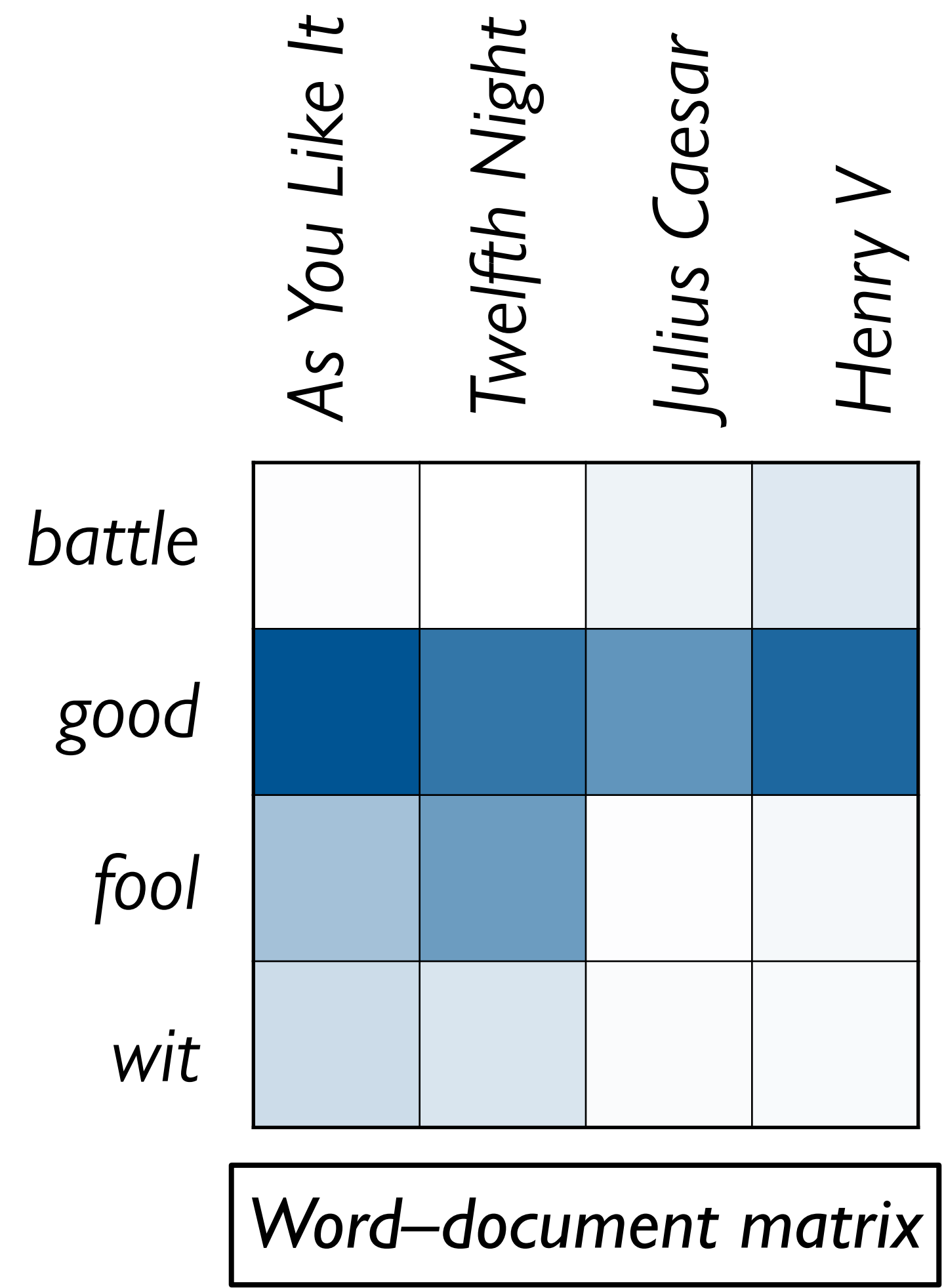
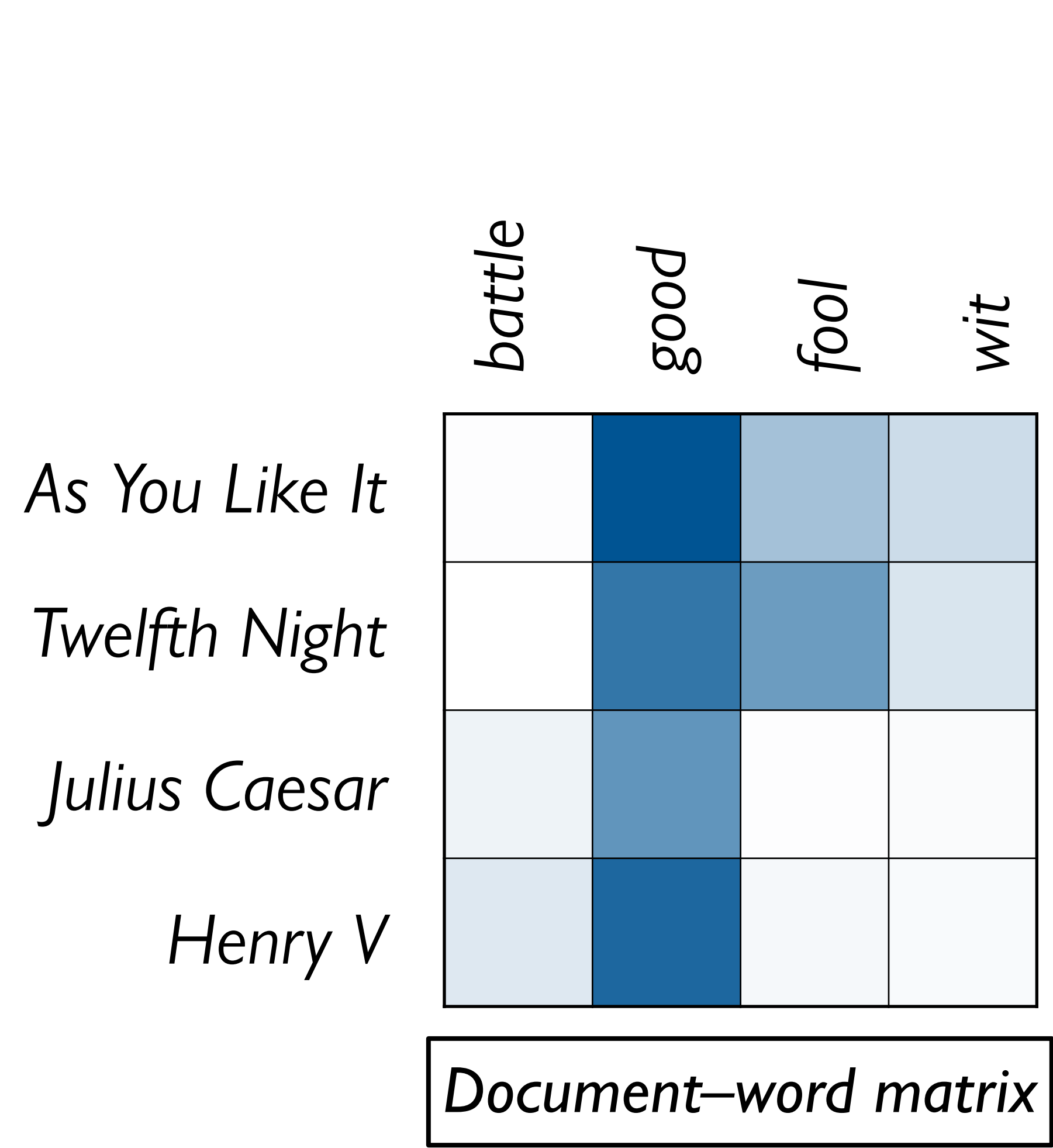
Word vectors

	<i>battle</i>	<i>good</i>	<i>fool</i>	<i>wit</i>
<i>As You Like It</i>				
<i>Twelfth Night</i>				
<i>Julius Caesar</i>				
<i>Henry V</i>				

Document–word matrix

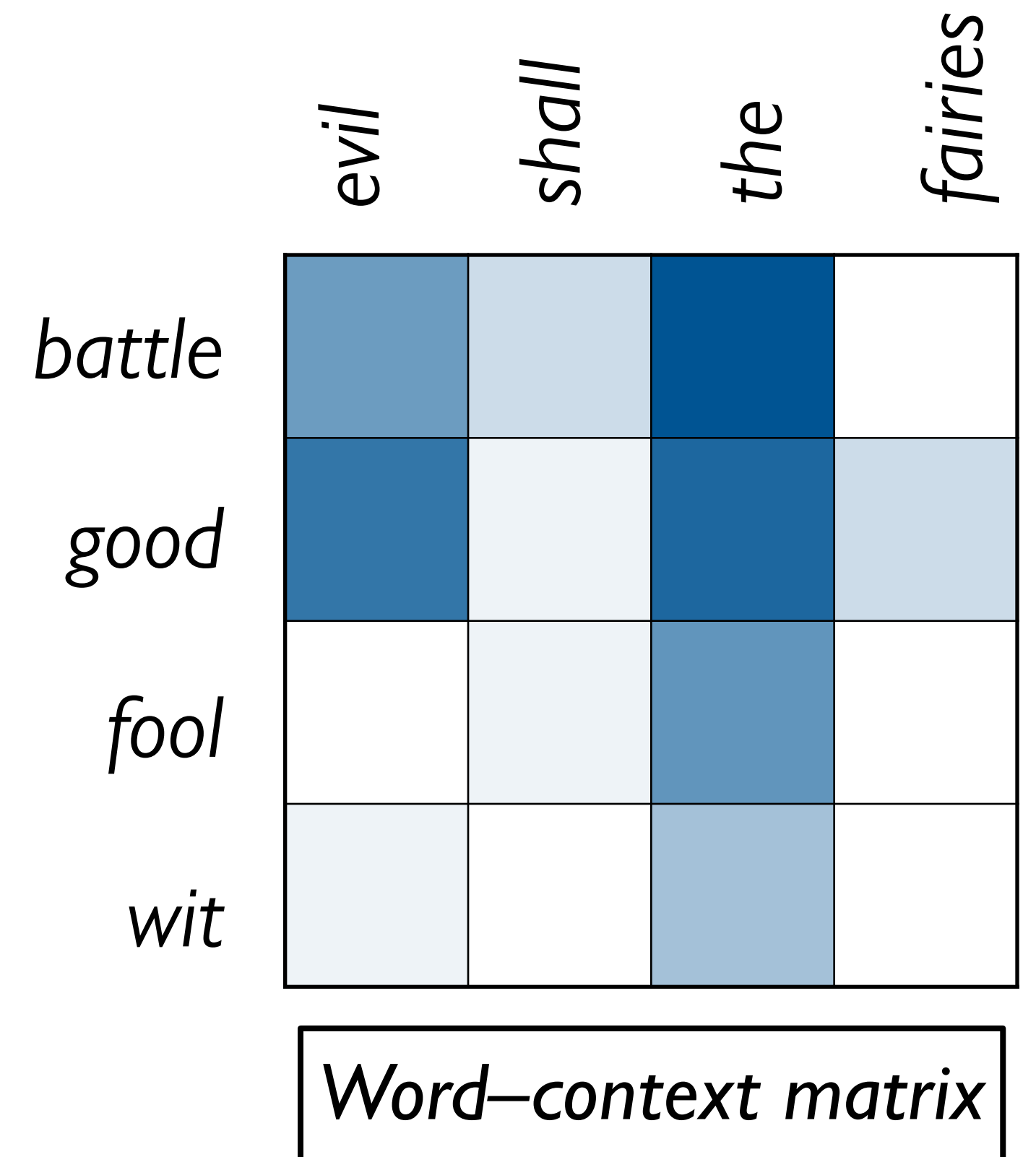
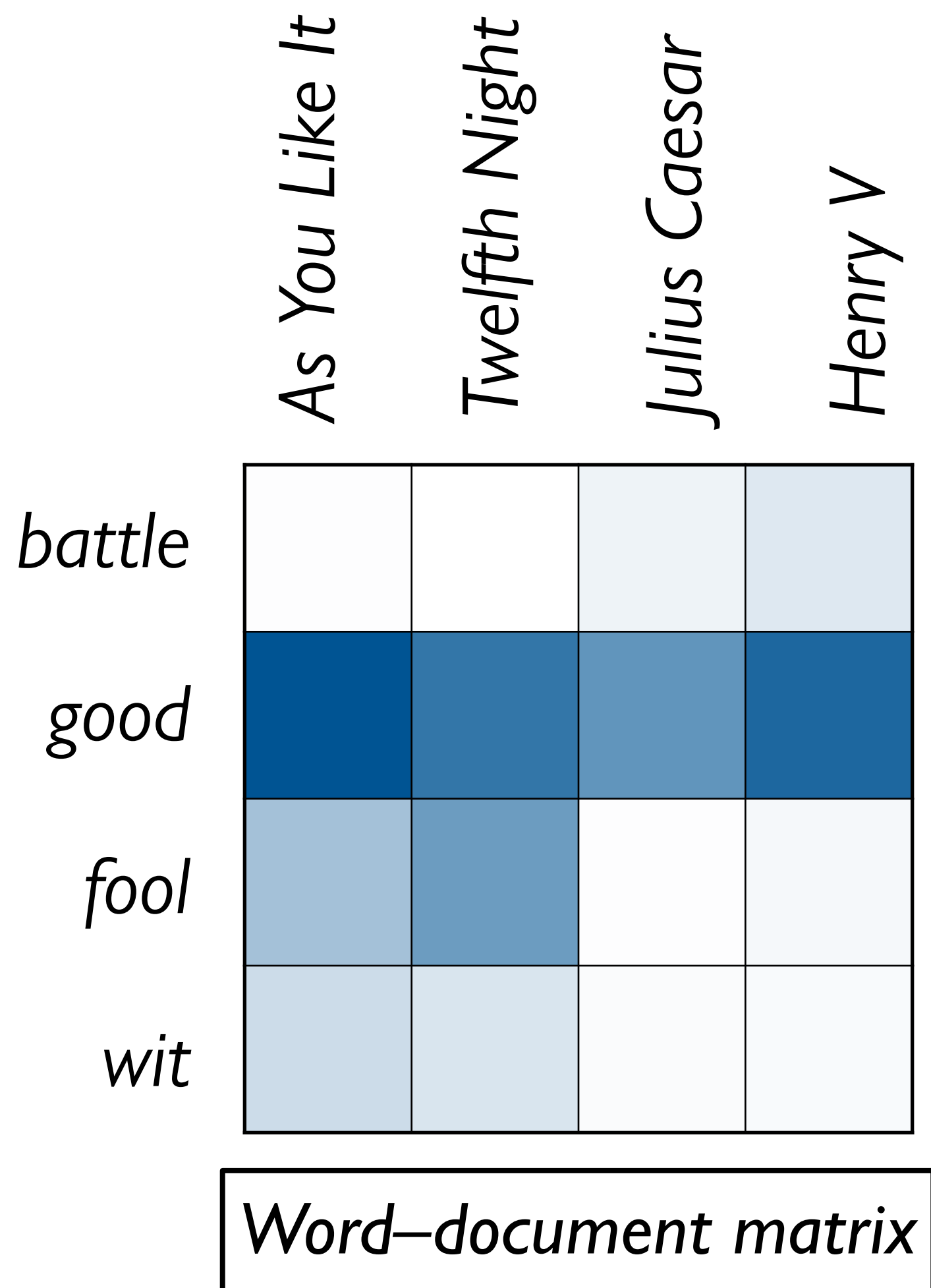
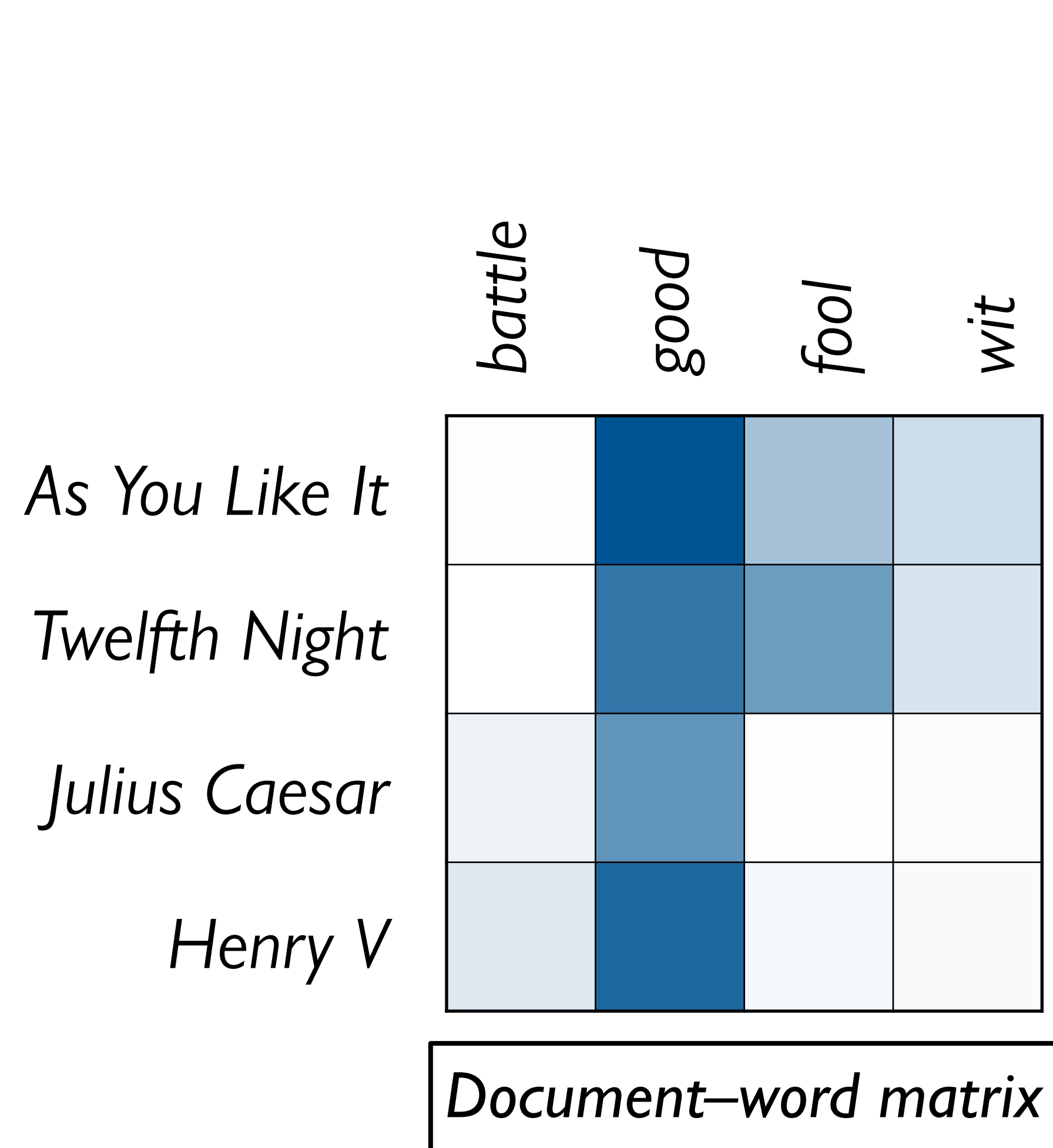
*A document can be represented
as the number of times each
word occurs in it.*

Word vectors

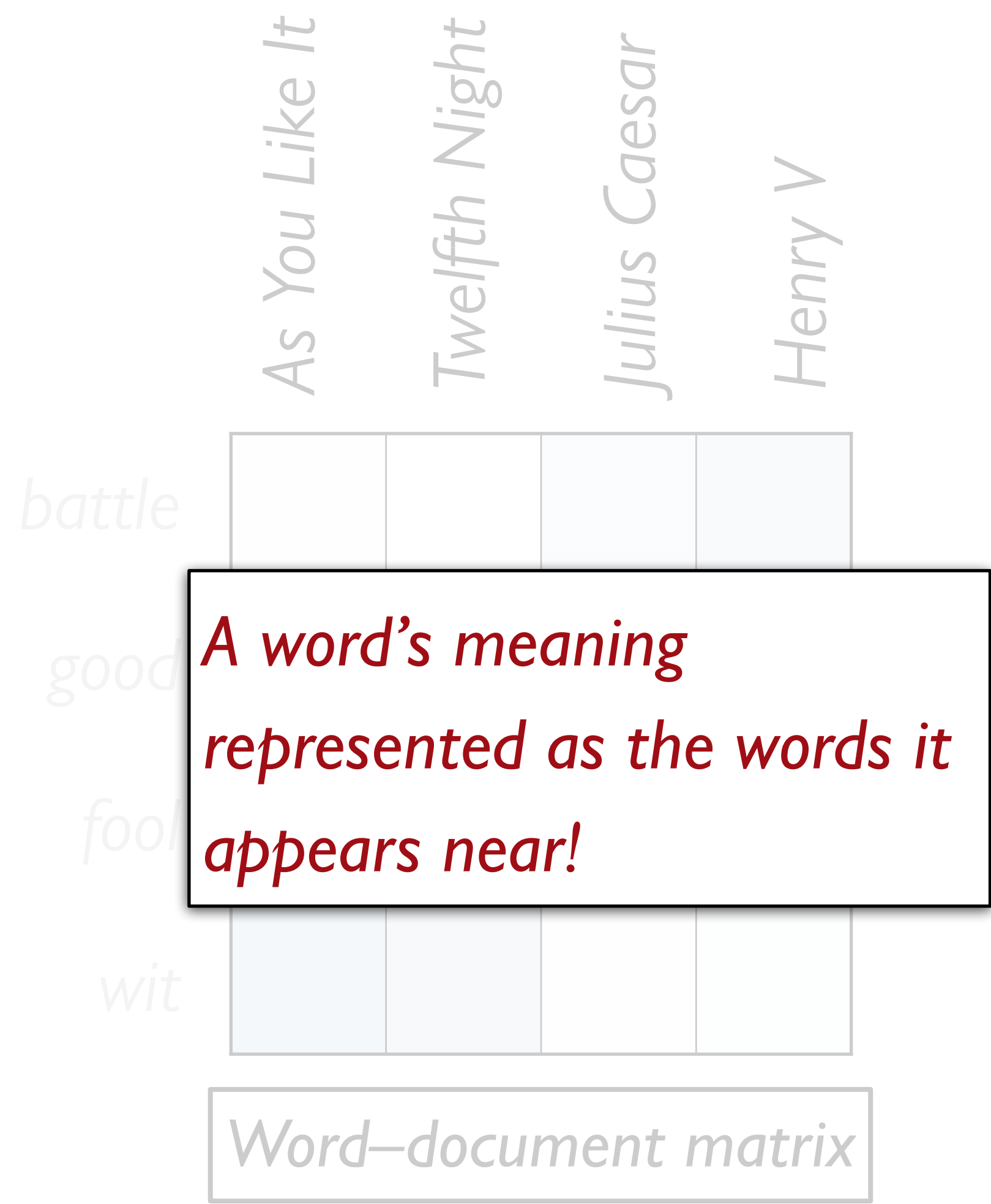
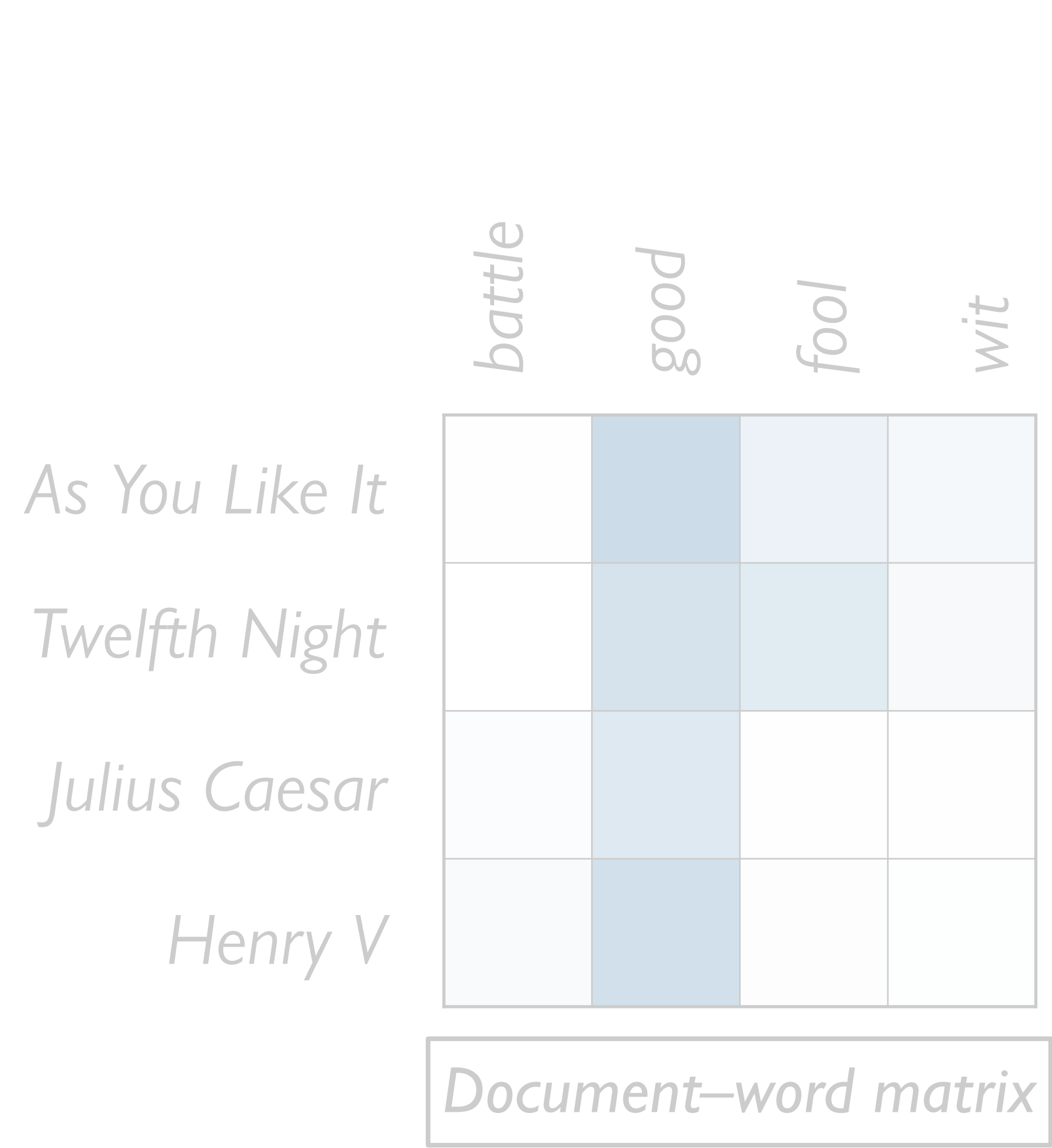


A word’s meaning can be represented as the documents it appears in!

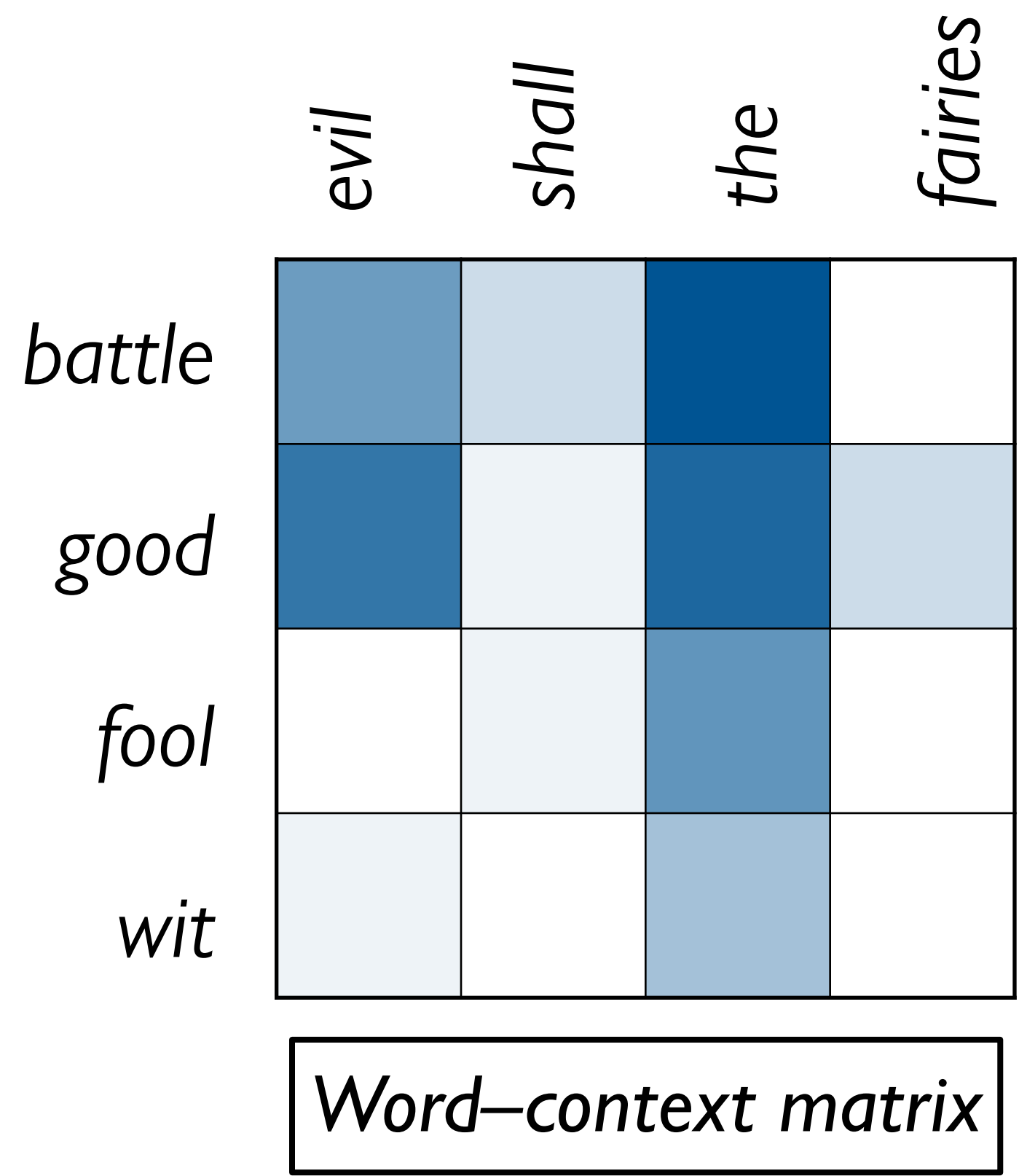
Word vectors



Word vectors



*A word’s meaning
represented as the words it
appears near!*



If a corpus can be split into short documents (e.g., tweets or headlines), you might use the entire document as the context.

Otherwise, we might choose a fixed *context window* of L tokens to either side, e.g., $L = 2$:

...lemon , a *tablespoon of apricot jam* , a pinch...

c_1 c_2 w c_3 c_4

target

*They picked up red **apples** that had fallen to the ground.*

*Eating **apples** is healthy.*

*She ate a red **apple**.*

*Pick an **apple**.*

They picked up red apples that had fallen to the ground.

Eating apples is healthy.

She ate a red apple.

Pick an apple.

They *picked up* red *apples* that had fallen to the ground.

Eating *apples* is healthy.

She ate a red *apple*.

Pick an *apple*.

	<i>a</i>	<i>be</i>	<i>eat</i>	<i>fall</i>	<i>have</i>	<i>healthy</i>	<i>pick</i>	<i>red</i>	<i>that</i>	<i>up</i>
<i>apple</i>	1	2	1	1	1	1	2	2	1	1

If we count context words in a larger corpus, we get counts like these:

	<i>eat</i>	<i>fall</i>	<i>ripe</i>	<i>slice</i>	<i>peel</i>	<i>tree</i>	<i>throw</i>	<i>fruit</i>	<i>pie</i>	<i>bite</i>	<i>crab</i>	...
<i>apple</i>	794	244	47	221	208	160	145	156	109	104	88	...

If we count context words in a larger corpus, we get counts like these:

	<i>eat</i>	<i>fall</i>	<i>ripe</i>	<i>slice</i>	<i>peel</i>	<i>tree</i>	<i>throw</i>	<i>fruit</i>	<i>pie</i>	<i>bite</i>	<i>crab</i>	...
<i>apple</i>	794	244	47	221	208	160	145	156	109	104	88	...
<i>orange</i>	265	22	25	62	220	64	74	111	4	4	8	...

If we count context words in a larger corpus, we get counts like these:

[illegible]

• • •

[illegible]

Given a word–context matrix, we can use each row as an embedding.

These embeddings are *sparse*, meaning they have a lot of zeros (when the word in the row never co-occurs with the word in the column).

What does this get us?

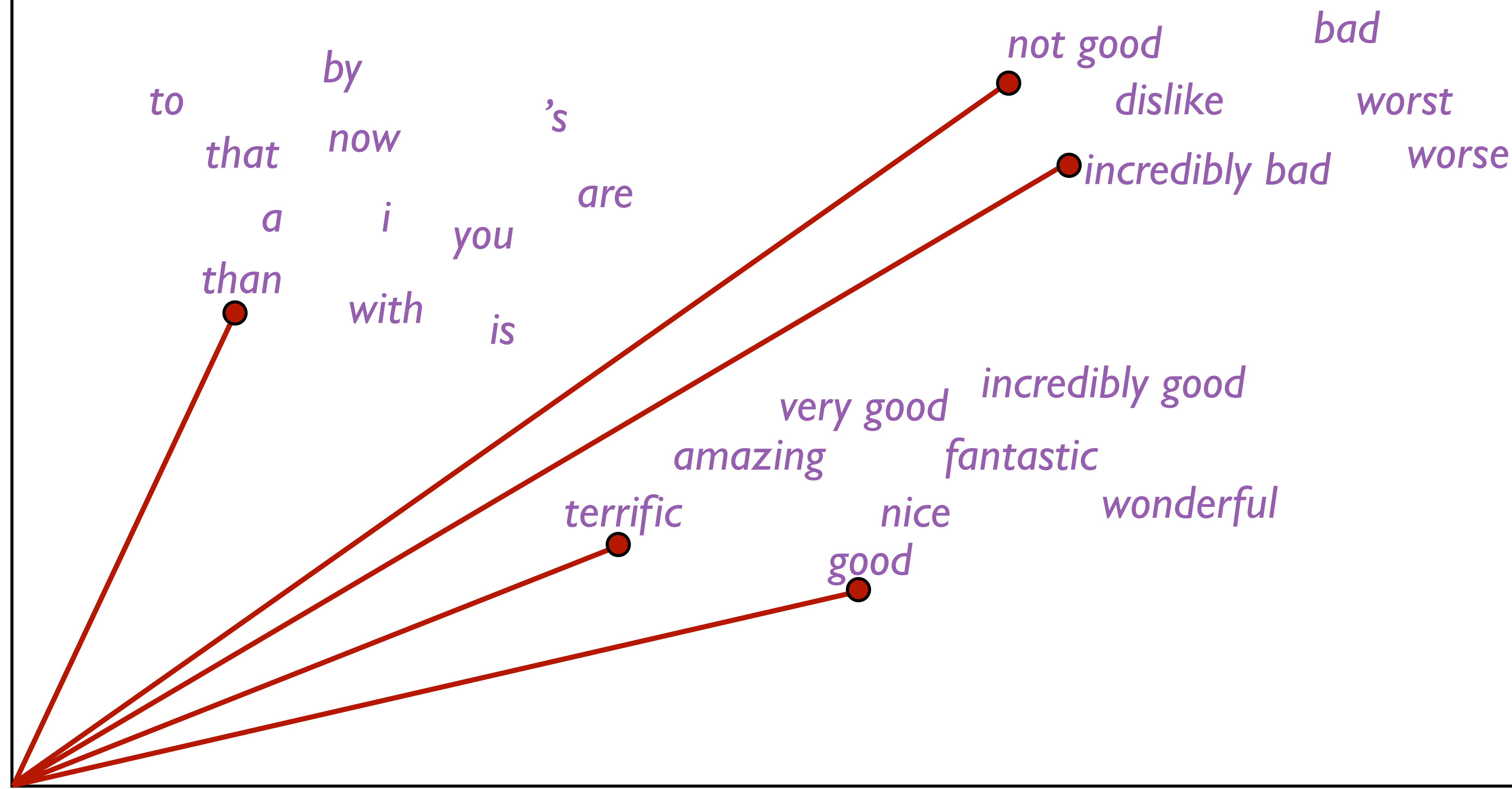
For tasks like sentiment analysis,

using words, we require the *same* word to be in training and test;

using embeddings, it's ok if *similar* words occurred!

Computing word similarity

Words used in the same context will cluster nearby in the same space.

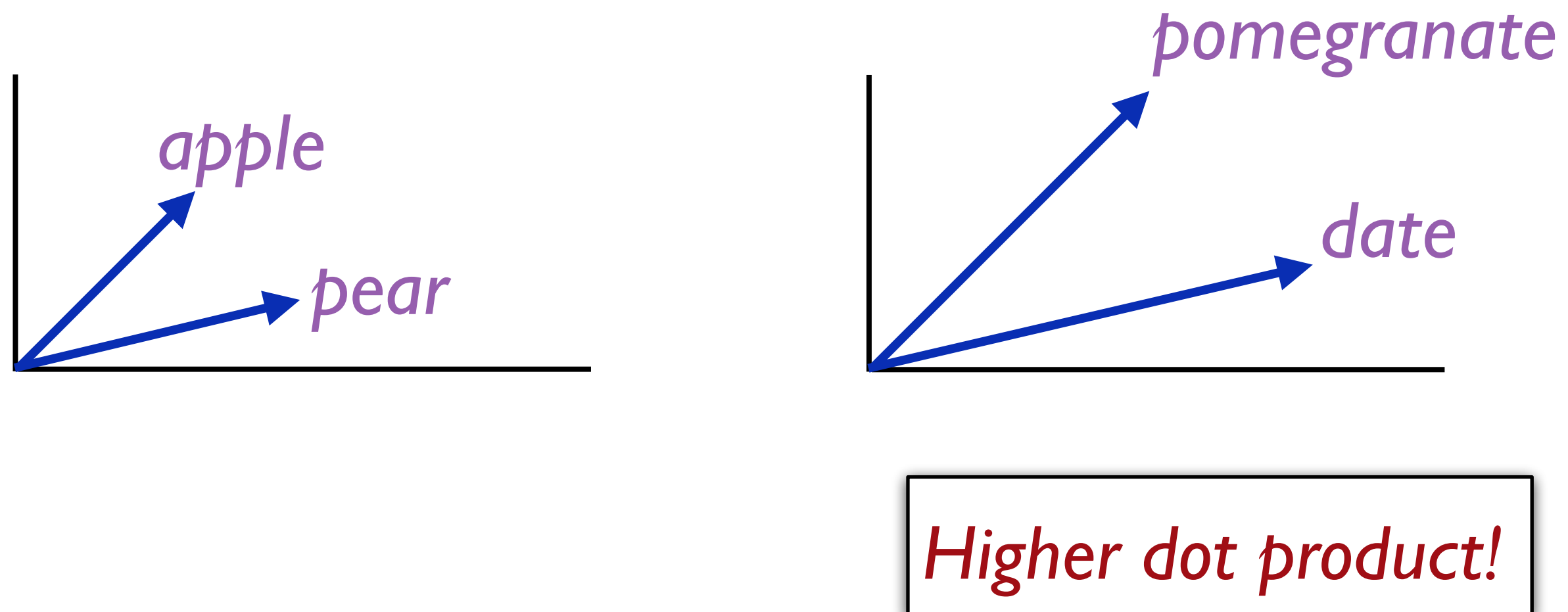


The dot product between two vectors

$$\mathbf{v} \cdot \mathbf{w} = \sum_{i=1}^N v_i w_i = v_1 w_1 + v_2 w_2 + \cdots v_N w_N$$

is a scalar that can be a useful similarity metric.

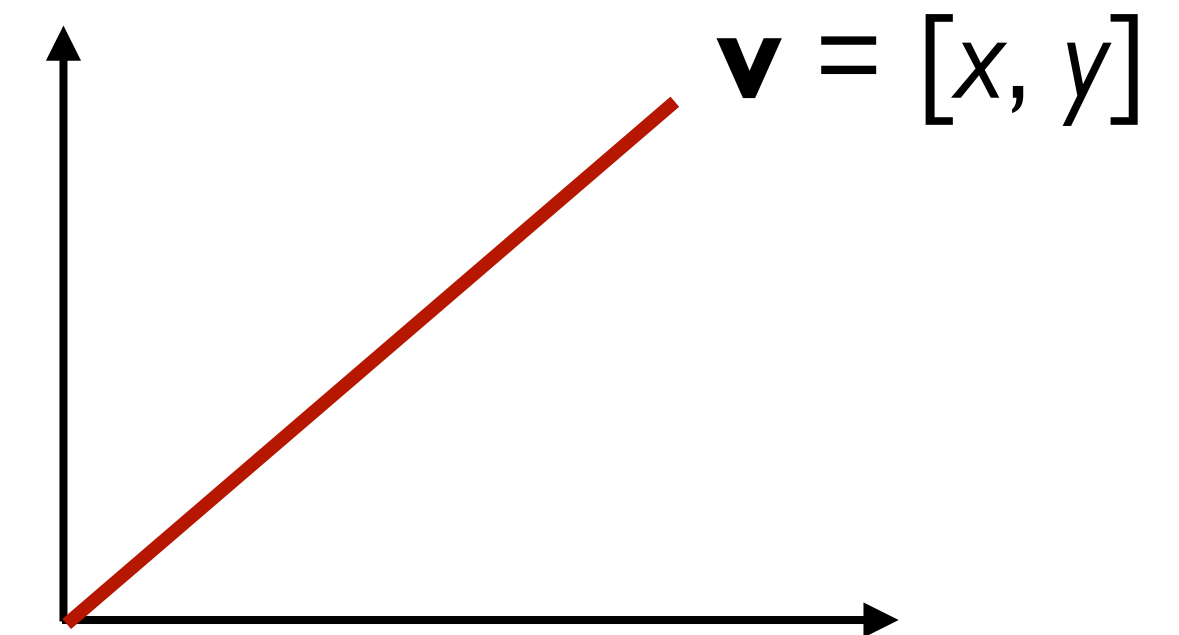
Problem: The longer the vectors are, the higher the dot product is



Solution: Normalize by vector length.

*For a 2-dimensional
vector*

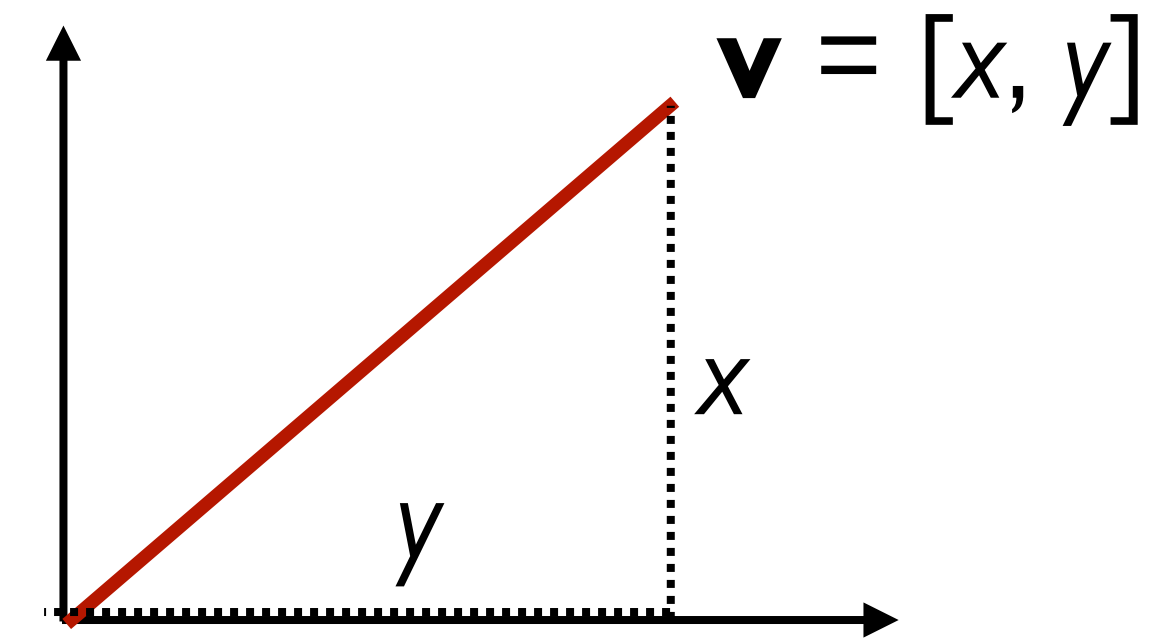
$$|\mathbf{v}| = \sqrt{x^2 + y^2}$$



Solution: Normalize by vector length.

*For a 2-dimensional
vector*

$$|\mathbf{v}| = \sqrt{x^2 + y^2}$$



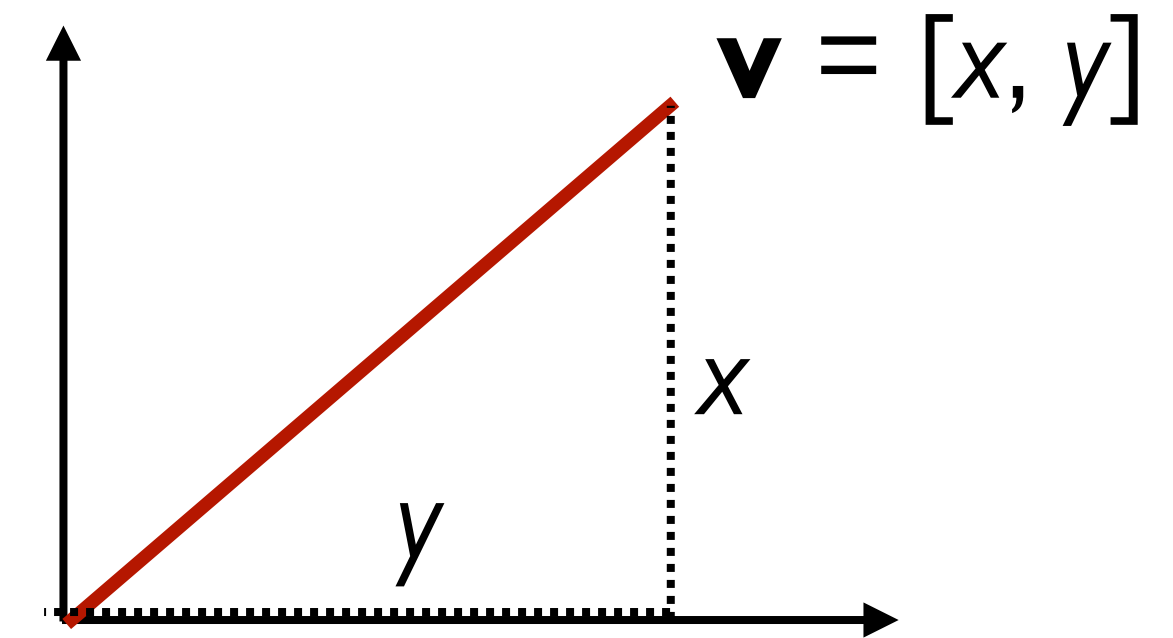
Solution: Normalize by vector length.

*For a 2-dimensional
vector*

$$|\mathbf{v}| = \sqrt{x^2 + y^2}$$

*For an N-dimensional
vector*

$$|\mathbf{v}| = \sqrt{\sum_{i=1}^N v_i^2}$$



Solution: Normalize by vector length.

$$|\mathbf{v}| = \sqrt{\sum_{i=1}^N v_i^2}$$
$$\frac{\mathbf{v} \cdot \mathbf{w}}{|\mathbf{v}| |\mathbf{w}|} = \frac{\sum_{i=1}^N v_i w_i}{\sqrt{\sum_{i=1}^N v_i^2} \sqrt{\sum_{i=1}^N w_i^2}}$$

Solution: Normalize by vector length.

$$|\mathbf{v}| = \sqrt{\sum_{i=1}^N v_i^2}$$

$$\cos(\mathbf{v}, \mathbf{w}) = \frac{\mathbf{v} \cdot \mathbf{w}}{|\mathbf{v}| |\mathbf{w}|} = \frac{\sum_{i=1}^N v_i w_i}{\sqrt{\sum_{i=1}^N v_i^2} \sqrt{\sum_{i=1}^N w_i^2}}$$

Surprise! It's just the cosine of the angle between the vectors!

Cosine as a similarity metric

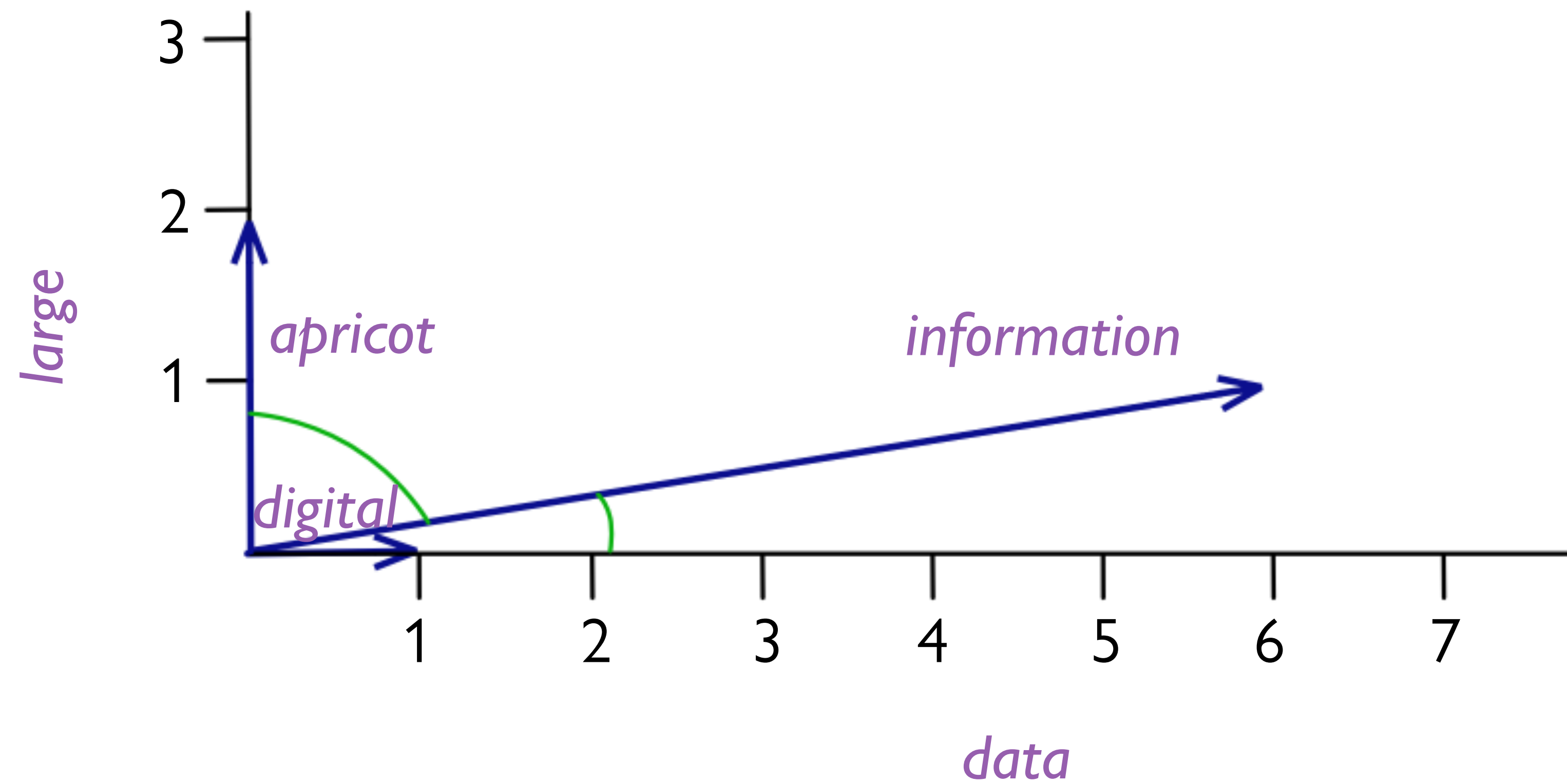
-1: Vectors point in opposite directions

+1: Vectors point the same direction

0: Vectors are orthogonal

But since raw frequency values are non-negative, the cosine for term-term matrix vectors ranges from 0–1.

Cosine as a similarity metric



Other similarity metrics are possible!

E.g., you could just use Euclidean distance:

$$\text{EucDist}(\mathbf{v}_1, \mathbf{v}_2) = \sqrt{(\mathbf{v}_1 - \mathbf{v}_2) \cdot (\mathbf{v}_1 - \mathbf{v}_2)}$$

But then you face the *curse of dimensionality*: in high dimensions, the Euclidean distance between pairs of vectors are all roughly the same.

Every modern NLP algorithm uses embeddings as the representation of word meaning.

But rather than simple count embeddings, which give us sparse vectors, they usually use embeddings learned with a classifier, which give us dense vectors.

We'll see the most popular technique for this – Word2vec – next class!

